Introduction to Monte Carlo Simulations

Bridging Asymptotic Theory and Finite Sample Properties

Minwoo Yoo

Ph.D. Candidate in Economics
George Washington University

Spring 2026

The Goal: Validating Inference

In econometrics, we rely on properties like Unbiasedness, Consistency, and Efficiency to trust our estimator $\hat{\theta}$.

Standard practice uses **Analytical Derivations** based on asymptotic theory.

Why Analytical Derivation is Not Enough

Analytical methods often fail in practical research settings:

1. **Finite Sample Bias:** Asymptotic approximations may be inaccurate when N is small.
2. **Intractability:** Closed-form solutions (integrals) may not exist for complex structural models.
3. **Complexity:** Deriving variance formulas for algorithmic estimators (e.g., multi-stage GMM) is often impossible.

Monte Carlo Simulation offers a computational alternative.

Scenario A: Finite Sample Bias

The Theoretical Context:

- Asymptotic theory ensures $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$.
- However, empirical research often deals with limited observations (e.g., $N = 30$).

The Problem:

- The finite-sample distribution may be skewed or have fat tails (Kurtosis ¿ 3), rendering standard t-tests invalid.

Scenario A: Finite Sample Bias

The Theoretical Context:

- Asymptotic theory ensures $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$.
- However, empirical research often deals with limited observations (e.g., $N = 30$).

The Problem:

- The finite-sample distribution may be skewed or have fat tails (Kurtosis ¿ 3), rendering standard t-tests invalid.

The Computational Breakthrough

Simulate the Finite Sample Distribution.

1. Define the Data Generating Process (DGP) with known θ.
2. Draw $R = 10,000$ samples of size $N = 30$.
3. Compute $\hat{\theta}_r$ for each replication.
4. **Analyze:** Plot the histogram to visually inspect convergence to Normality.

Scenario B: Intractable Integrals in Structural Models

The Context (e.g., BLP Demand Estimation):

- Calculating market shares or expected utilities often requires integrating over unobserved heterogeneity:

$$\sigma_j(p) = \int P_j(p, \nu) f(\nu) d\nu$$

The Obstacle:

- High-dimensional integrals often lack closed-form analytical solutions.

Scenario B: Intractable Integrals in Structural Models

The Context (e.g., BLP Demand Estimation):

- Calculating market shares or expected utilities often requires integrating over unobserved heterogeneity:

$$\sigma_j(p) = \int P_j(p, \nu) f(\nu) d\nu$$

The Obstacle:

- High-dimensional integrals often lack closed-form analytical solutions.

Monte Carlo Integration

Applying the Law of Large Numbers (LLN). We replace the integral with a simulation-based average:

$$\int g(\nu) f(\nu) d\nu \approx \frac{1}{R} \sum_{r=1}^{R} g(\nu_r)$$

By drawing random draws from $f(\nu)$, we transform calculus into arithmetic.

Scenario C: Complex "Black Box" Estimators

The Context:

- Some estimators rely on complex optimization routines (e.g., complex GMM or multi-stage estimators) where deriving the analytical variance-covariance matrix is computationally prohibitive.

The Question:

- How do we report Standard Errors without an explicit formula?

Scenario C: Complex "Black Box" Estimators

The Context:

- Some estimators rely on complex optimization routines (e.g., complex GMM or multi-stage estimators) where deriving the analytical variance-covariance matrix is computationally prohibitive.

The Question:

- How do we report Standard Errors without an explicit formula?

Empirical Variance Approach

Simulation-Based Inference.

- Replicate the estimation process R times on simulated data generated from the assumed model.
- Compute the empirical variance of the resulting estimates:

$$\widehat{Var}(\hat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\hat{\theta}})^2$$

Experiment Design: Testing the Central Limit Theorem

Objective: Test the validity of the Central Limit Theorem (CLT) for the sample mean $\bar{Y}$ when $N = 25$.

Simulation Blueprint: We control the Data Generating Process (DGP) to verify the theory.

1. **Assumption:** True population follows $Y \sim N(0, 1)$.
2. **Action:** Draw $N = 25$ random observations.
3. **Estimation:** Calculate the sample mean $\bar{Y}$.
4. **Repetition:** Repeat 100,000 times (Vectorized implementation).
5. **Verification:** Compare the empirical distribution against the theoretical Normal distribution.

Implementation: Vectorized Simulation in Python

The following code demonstrates a vectorized implementation for computational efficiency.

```python
import numpy as np
import matplotlib.pyplot as plt

# 1. DGP Configuration
N_SAMPLES = 25              # Finite sample size
TRUE_MEAN = 0               # Parameter theta (truth)
TRUE_STD = 1                # Population standard deviation
MC_REPS = 100000            # Number of Monte Carlo replications

# 2. Vectorized Simulation (Efficient)
# Generate (MC_REPS x N_SAMPLES) matrix from N(0,1)
# This avoids slow Python loops
data_matrix = np.random.normal(TRUE_MEAN, TRUE_STD, (MC_REPS, N_SAMPLES))

# 3. Estimator Calculation
# Compute sample mean across the rows (axis=1)
sample_means = np.mean(data_matrix, axis=1)
```

Visualization: Sampling Distribution

Visualizing the results to assess asymptotic approximation.

```
# 5. Plotting the Histogram
plt.figure(figsize=(10, 6))
plt.hist(sample_means, bins=50, color='#B03A2E', edgecolor='white', alpha=0.8)

plt.axvline(TRUE_MEAN, color='black', linestyle='dashed', linewidth=1.5, label='True Mean')
plt.title(f"Finite Sample Distribution of Mean (N={N_SAMPLES})", fontsize=14)
plt.xlabel("Estimated Parameter Value")
plt.ylabel("Frequency")
plt.legend()

plt.show()
```

Comparison: Monte Carlo vs. Bootstrap

	Monte Carlo Simulation	**Bootstrapping**
Assumption	DGP is **known** (User-defined).	DGP is **unknown** (Empirical Distribution).
Usage	Validating estimators, power analysis, checking theoretical properties.	Inference (SEs, CIs) for a specific dataset when formulas are complex.

Conclusion: Monte Carlo methods allow economists to verify the reliability of estimators when analytical derivations are intractable or when asymptotic assumptions are violated.